Scaling laws for the movement of people between locations in a large city

G. Chowell[1,2], J. M. Hyman[2], S. Eubank[2], C. Castillo-Chavez[1,2]

MS B258

[1] Los Alamos National Laboratory

Los Alamos, NM 87545, U.S.A.

[2] Department of Biological Statistics and Computational Biology, Cornell University,

Warren Hall, Ithaca, NY 14853-7801, U.S.A *

Abstract

Large scale simulations of the movements of people in a "virtual" city and their analyses are used to generate new insights into understanding the dynamic processes that depend on the interactions between people. Models, based on these interactions, can be used in optimizing traffic flow, slowing the spread of infectious diseases or predicting the change in cell phone usage in a disaster. We analyzed the data generated from the simulated movements of 1.6 million individuals in a computer (pseudo agent-based) model for Portland, OR. This city is mapped into a graph with 181,205 nodes representing physical locations such as buildings. Connecting edges model individual's flow between nodes. Edge weights are constructed from the daily traffic of individuals moving between locations. The number of edges leaving a node (out-degree), the edge weights (out-traffic), and the edge-weights per location (total out-traffic) are fitted well by power law distributions. The power law distributions also fit subgraphs based on work, school, and social/recreational activities. The resulting weighted graph is a "small world" and has scaling laws consistent with an underlying hierarchical structure. We also explore the time evolution of the largest connected component and the distribution of the component sizes. We observe a strong linear correlation between the out-degree and total out-traffic distributions and significant levels of clustering. We discuss how these network features can be used to characterize social networks and their relationship to dynamic processes.

*Los Alamos Unclassified Report LA-UR-02-6658.

1 Introduction

Similar scaling laws and patterns have been detected in networks of scientific collaboration [1][2][3], cellular networks [4][5], the Internet [6], and the *World Wide Web* [7][8]. These networks exhibit the "small world effect," [9][10] where the average number of edges needed to connect *any* pair of nodes is small and high clustering is observed, a characteristic absent in random networks [11].

The connectivity distribution of many networks has been captured by power-law distributions, $P(k) \propto k^{-\gamma}$, where the exponent γ characterizes the underlying scaling of the network and k denotes the node's degree or incidence.

Barabási and Albert (BA) introduced an algorithm capable of generating networks with a power-law connectivity distribution ($\gamma = 3$). The BA algorithm generates networks where nodes connect, with higher probability, to nodes that have a accumulated higher number of connections and stochastically generates networks with a power-law connectivity distributions ($P(k) \propto k^{-\gamma}$), in the appropriate scale. We generate a directed graph from the simulated movement of 1.6 million individuals *in* or *out* of 181, 205 locations in Portland, OR. The 181, 205 nodes represent locations in the city and the edges connections between nodes. The edges are weighted by daily traffic (movement of individuals) *in* or *out* of these locations. The statistical analysis of the network topology reveals that it is a small world with power-law decay in the out-degree distribution of locations (nodes). The resulting network has scaling laws consistent with an underlying hierarhical structure [12, 13]. The out-traffic (weight of the full network) and the total out-traffic (total weight of the out edges per node) distributions are also fitted to power laws. We show that the joint distribution of the out-degree and total out-traffic distributions decays linearly in an appropriate scale. We also explore the time evolution of the largest component and the distribution of the component sizes.

2 Location-based network

A "typical" realization by the Transportation Analysis Simulation System (TRANSIMS) simulates the dynamics of 1.6 million individuals in the city of Portland as a directed network, where the nodes represent locations (i.e. buildings, households, schools, etc.) and the directed edges (between the nodes) represent the movement (traffic) of individuals between locations (nodes). Traffic intensity is modeled by the nonsymmetric mobility matrix $W = (w_{ij})$ of traffic weights assigned to all directed edges in the network ($w_{ij} = 0$ means that there is no directed

edge connecting node i *to* node j).

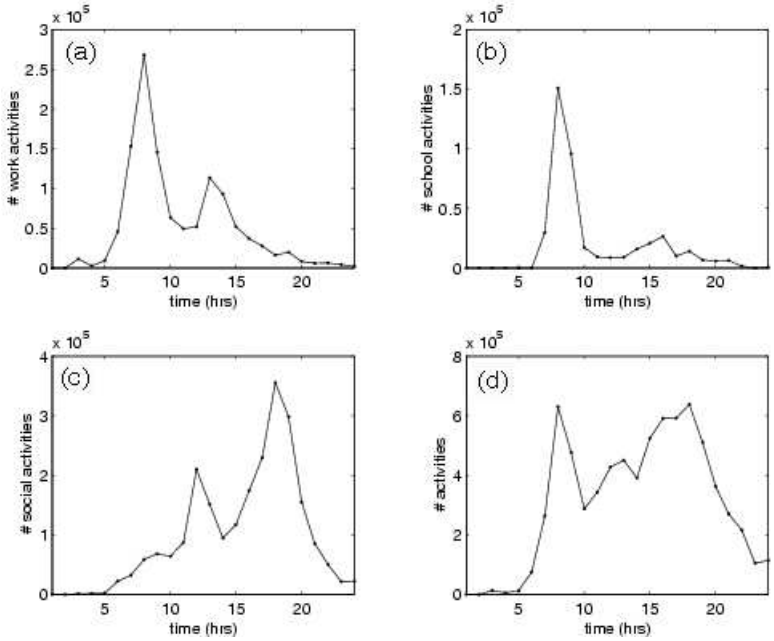

Figure 1: Activity is the movement of an individual to the location where the activity will be carried out. (a)work activities, (b)school activities, (d)social activities, and (d)total number of individual activities as a function of time (hours) of a 'typical' day in the city of Portland, OR.

TRANSIMS [14] is a pseudo agent-based simulation model of the movement of individuals in virtual regions or cities. TRANSIMS simulates the movement of individuals in this virtual region through the city's transportation network (mass or car transportation). First, a detailed representation of the region is created, mobility information for each individual is incorporated from survey data and the known transportation infrastructure is used to connect the sites. The simulation is tunned so that movement data, obtained from transportation planning surveys of detailed information on people's movement (daily trips), is accurately captured. Data also include information on activity types (see Figure 1), origins, destinations, routes, timing, and forms of transportation used. TRANSIMS calculates the simulated movements of 1.6 million individuals in a typical day [14]. The simulated Portland data set includes the time at which each individual leaves a location and the time of arrival to its next destination (node). These data are used to calculate the average number of people at each location and the traffic between any two locations on a typical day. (Table 1 shows a sample of a Portland activity file generated by TRANSIMS).

TRANSIMS' simulation of the network of social interactions is based on the assumption

that the locations of people's activity choices is constrained by the transportation infrastructure. The synthetic population is endowed with matching demographic characteristics derived from data (joint distributions from census data). Observations made on the daily activity patterns of several thousand households (survey data) are used as templates for the modeling of synthetic households with matching demographics. Locations where activities are carried out are estimated from observed land use patterns, travel times and costs of transportation alternatives. These locations are fed into a routing algorithm that finds the minimum cost paths that are consistent with individual choices [15, 16, 17]. The simulation resolution is of 7.5 meters and 1 second. The simulator provides an updated estimate of time-dependent travel times for each edge in the network, including the effects of congestion, to the Router and location estimation algorithms, which generate traveling plans. Since the entire process estimates the demand on a transportation network from census data, land use data, and activity surveys, these estimates can thus be applied to assess the effects of hypothetical changes such as building new infrastructures or changing downtown parking prices. Methods based on observed demand cannot handle such situations, since they have no information on what generates the demand. Simulated traffic patterns compare well to observed traffic and, consequently, TRANSIMS provides a useful planning tool.

Until recently, it has been difficult to obtain useful estimates on the structure of social networks. Certain classes of random graphs (scale-free networks [18], small-world networks [10, 19], or Erdos-Renyi random graphs [11, 20]), have been postulated as good representatives. In addition, data based models while useful are limited since they have naturally focused on small scales [21]. While most studies on the analysis of real networks are based on a single snapshot of the system, TRANSIMS provides powerful time dependent data of the evolution of a location-based network. Though social network (mobility) estimates are not strictly speaking part of TRANSIMS, the level of detail generated in the simulations is such that the aggregation of the movement of individuals at multiple temporal and spatial scales reveals important trends.

Table 1. Sample section of a TRANSIMS activity file. In this example, person 115 arrives for a social recreational activity at location 33005 at 19.25 o'clock and departs at 21.00 o'clock.

Person ID	Location ID	Arrival time(hrs)	Departure time(hrs)	Activity type
115	4225	0.0000	7.00	home
115	49296	8.00	11.00	work
115	21677	11.2	13.00	work
115	49296	13.2	17.00	work
115	4225	18.00	19.00	home
115	33005	19.25	21.00	social/rec
115	4225	21.3	7.00	home
220	8200	0.0000	8.50	home
220	10917	9.00	14.00	school
220	8200	14.5	18.00	home
220	3480	18.2	20.00	soc/rec
220	8200	20.3	8.6	home

3 Power law distributions

We calculate the statistical properties of a typical day in the location-based network of this vitual city from mobility data generated by TRANSIMS (see Table 2).

The *average out-degree*, $\bar{k}$, is $\bar{k} = \sum_{i=1}^{n} k_i/n$ where k_i is the degree for node i and n is the total number of nodes in the network. For the portland network $\bar{k} = 29.88$ and the *out-degree distribution* exhibits power law decay with scaling exponent ($\gamma \approx 2.7$). The *out-traffic* (edge weights) and the *total out-traffic* (edge-weights per node) distributions are also fitted well by power laws. The *average distance* between nodes L is defined as the median of the means L_i of the shortest path lengths connecting a vertex $i \in V(G)$ to all other vertices [22]. For our network, $L = 3.38$, which is small when compared to the size of the network. In fact, the *diameter* (D) of the graph (the largest of all possible shortest paths between all the locations) is only 9. L and D are measured using a breadth first search (BFS) algorithm [23] on a randomly selected subgraph of size $90,000$ ($\approx 50\%$ the size of the whole network) ignoring the edge directions. The *clustering coefficient*, C, quantifies the extent to which neighbors of a node are also neighbors of each other [22]. The clustering coefficient of node i, C_i, is given by

$$C_i = |E(\Gamma_i)| \,/ \binom{k_i}{2}$$

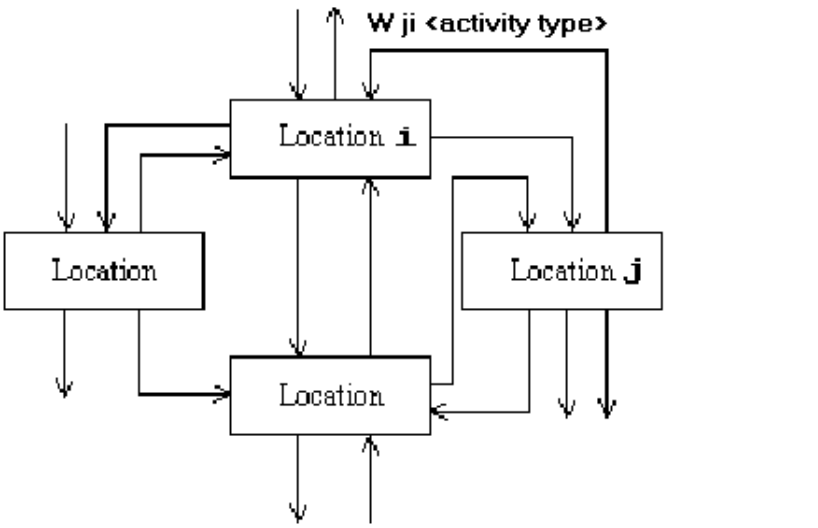

Figure 2: The nodes represent locations connected via directed edges based on the traffic or movement of individuals (activities) between the locations. The weights (w_{ij}) represent the daily traffic from location i to location j.

where $|E(\Gamma_i)|$ is the number of edges in the neighborhood of i (edges connecting the neighbors of i not including i itself) and $\binom{k_i}{2}$ is the maximal number of edges that could be drawn among the k_i neighbors of node i. The clustering coefficient C of the whole network is $C = \sum_{i=1}^{n} C_i/n$. For a typical *random graph* with $181,205$ nodes and average connectivity $\bar{k} = 29.88$, the clustering coefficient $C \approx \bar{k}/n \approx 0.000166$ [22]. The clustering coefficient for our location-based network, ignoring edge directions, is $C = 0.0584$, which is roughly 350 times larger than C_{rand}. Highly clustered networks have been observed in other systems [9] including the electric power grid of western US. This grid has a clustering coefficient $C = 0.08$, about 160 times larger than the expected value for an equivalent random graph [22]. The few degrees of separation between the locations of the (highly clustered) network of the city of Portland "make" it a small world [22, 10, 19].

Table 2. Statistical properties of the location-based network for Portland. The clustering coefficient seems "small" but it is roughly 350 times larger than the expected clustering coefficient for an equivalent random graph of the same size n and average degree $\bar{k}$. The small average distance between nodes L and the significant levels of clustering C make this network a "small world."

Statistical properties	Value
Total nodes (n)	181205
Size of the giant component	181192
Total directed edges (m)	5416005
Average degree ($\bar{k}$)	29.88
Clustering coefficient (C)	0.0584
Average distance between nodes (L)	3.38
Diameter (D)	9.0

Many real-world networks exhibit properties that are consistent with underlying hierarhical organizations. These networks have groups of nodes that are highly interconnected with few or no edges connected to nodes outside their group. Hierarchical structures of this type have been characterized by the clustering coefficient function $C(k)$, where k is the node degree. A network of movie actors, the semantic web, the *World Wide Web*, the Internet (autonomous system level), and some metabolic networks [12, 13] have clustering coefficients that scale as k^{-1}.

The clustering coefficient as a function of degree (ignoring edge directions) in the Portland network exhibits similar scaling at various levels of aggregation that include, the whole network and subnetworks constructed by activity type (work, school and social/recreational activities, see Figure 3). We constructed subgraphs based on activity types. The clustering coefficient of the subnetworks generated from work, school, and social/recreational activities are: 0.0571, 0.0557, and 0.0575, respectively. The largest clustering coefficient and closest to the overall clustering coefficient ($C = 0.0584$) correponds to the subnetwork constructed from social/recreational activities. It seems that the whole network, as well as the selected activity subnetworks, support a hierarchical structure albeit the nature of such structure (if we choose to characterize by the power law exponent) is not universal. This agrees with relevant theory [13].

Understanding the temporal properties of networks is critical to the study of superimposed dynamics such as the spread of epidemics on networks. Most studies of superimposed processes on networks assumes that the contact structure is fixed (see for example [24, 25, 26, 27, 28, 29,

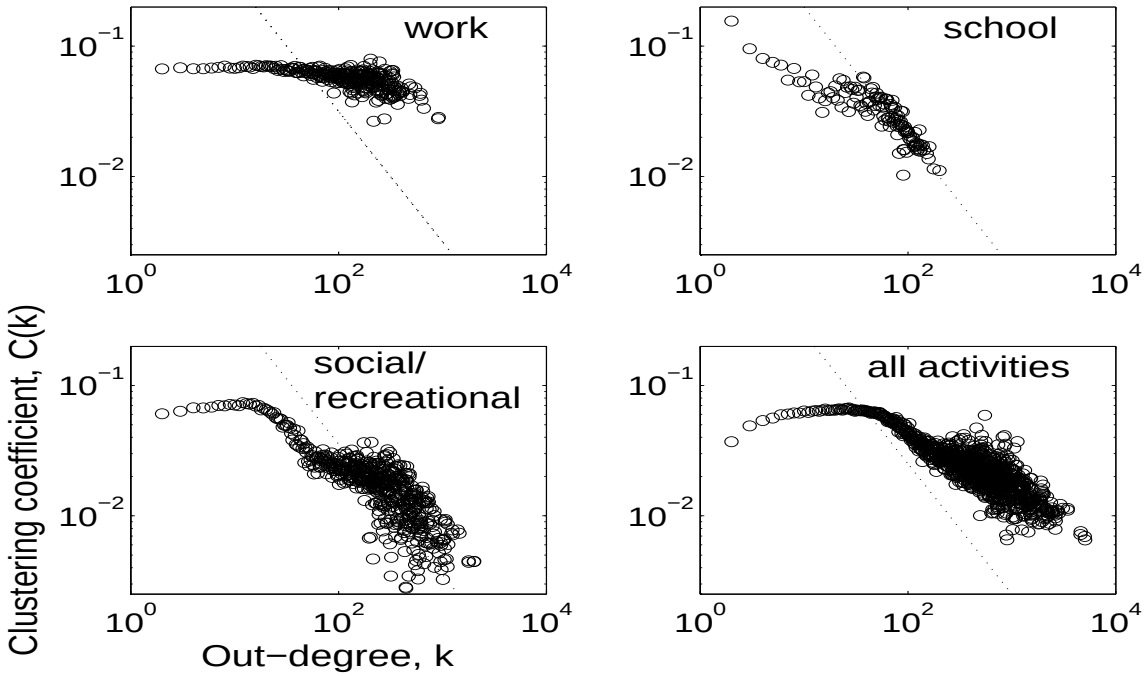

Figure 3: Log-log plots of the clustering coefficient as a function of the out-degree for subnetworks constructed from work activities, school activities, social activities, and all the activities. The dotted line has slope -1. Notice the scaling k^{-1} for the school and social/recreational activities. However, for the subnetwork constructed from work activities, the clustering coefficient is almost independent of the out-degree k.

30, 31, 32]). Here, we take a look at the time evolution of the largest connected component of the location-based network of the city of Portland. We observe that the network undergoes a sharp transition at approximately 6 a.m. (see Fig. 4) in the morning at which a 'giant component' appears. The distribution of the sizes of the components (clusters of locations) follows a power law that gets steeper in time until it dissolves as the giant componet forms (see Figure 5).

To identify the relevance of the temporal trends, we computed the out-degree distribution of the network for three different time intervals: The morning from 6 a.m to 12 p.m.; the workday from 6 a.m. to 6 p.m.; and the full 24 hours. In the morning phase, the out-degree distribution has a tail that decays as a power law with $\gamma \simeq 3$ (for the workday $\gamma \simeq 2.6$ and for the full day $\gamma \simeq 2.7$). The distribution of the out-degree data has two scaling regions: the number of locations is approximately constant for out-degree $k < 20$ and then decays as a power law for high degree nodes (Fig. 6).

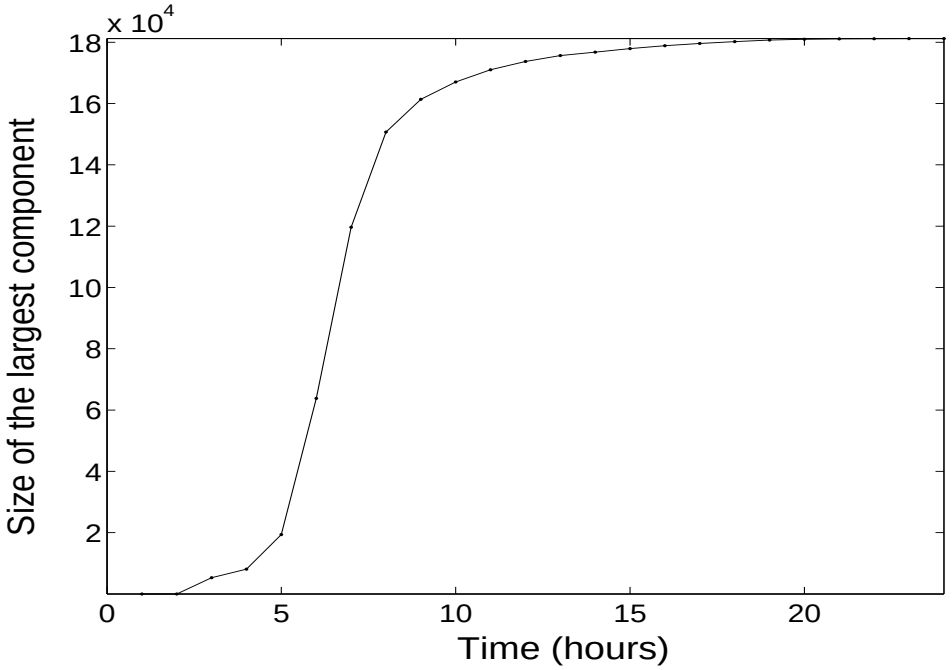

Figure 4: The size of the largest component (cluster) over time. A sharp transition is observed at about 6 a.m when people move from home to work or school.

The strength of the connections in the location-based network is measured by the traffic (flow of individuals) between locations in a "typical" day of the city of Portland. The log-log plot of the out-traffic distributions for three different periods of time (Fig. 7) exhibits power law decay with exponents, $\gamma \simeq 3.56$ for the morning, $\gamma \simeq 3.74$ for the workday, and $\gamma \simeq 3.76$ for the full day. The out-traffic distribution is characterized by a power law distribution for all values of the traffic-weight matrix W. This is not the case for the out-degree distribution of the network (see Figure 6) where a power law fits well only for sufficiently large degree k ($k > 10$).

The distribution of the total out-traffic per location, w_i's ($w_i = \sum_j w_{i,j}$), is characterized by two scaling regions. The tail of this distribution decays as a power law with exponent $\gamma = 2.74$ (Fig. 8). This is almost the same decay as the out-degree distribution ($\gamma = 2.7$) because the out-degree and the total out-traffic are highly correlated (with correlation coefficient $\rho = 0.94$).

4 Correlation between out-degree and total out-traffic

The degree of correlation between various network properties depend on the social dynamics of the population. The systematic generation and resulting structure of these networks is im-

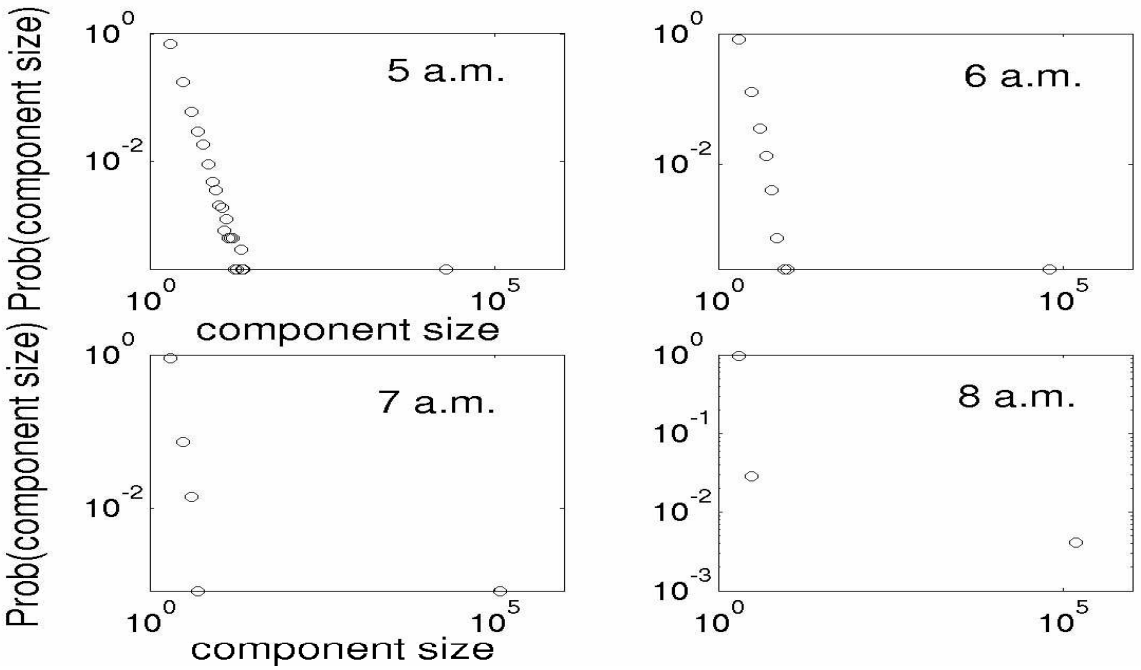

Figure 5: Log-log plot of the distribution of the component sizes at different times of the day.

portant to understand dynamic processes such as epidemics that "move" on these networks. Understanding the mechanisms behind these correlations will be useful in modeling fidelity networks.

In the Portland network, the out-degree k and total out-traffic v have a correlation coefficient $\rho = 0.94$ on a log-log scale with 95% of the nodes (locations) having out-degree and total out-traffic less than 100 (Fig. 9). That is, the density of their joint distribution $F(k, v)$ is highly concentrated near small values of the out-degree and total out-traffic distributions. The joint distribution supports a surface that decays linearly when the density is in log_e scale (Figure 10).

5 Conclusions

Strikingly similar patterns on data from the movement of 1.6 million individuals in a "typical" day in the city of Portland have been identified at multiple temporal scales and various levels of aggregation. The analysis is based on the mapping of people's movement on a weighted directed graph where nodes correspond to physical locations and where directed edges, connecting the

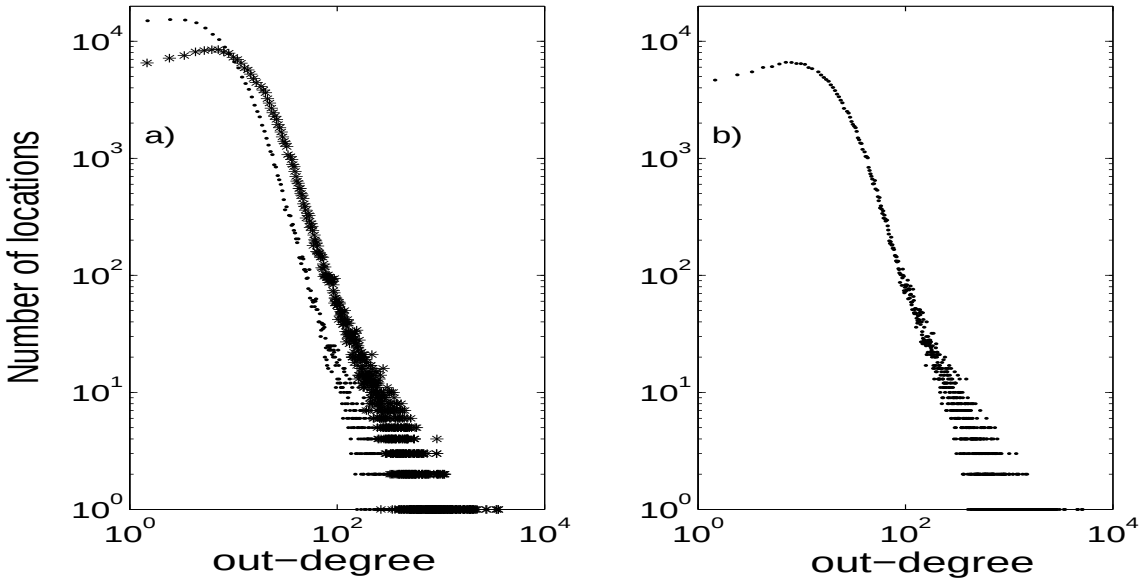

Figure 6: There are two distinct scaling regions for the number of nodes as a function of the out-degree of the nodes. There are approximately the same number of nodes with out-degree $k = 1, 2, ...10$ For $k > 0$ the number of nodes with a given out-degree decays as a power law $P(k) \propto k^{-\gamma}$ with $\gamma \simeq 3$ for the morning (6 a.m.-12 p.m.), $\gamma \simeq 2.6$ for the workday (6 a.m.-6 p.m.) and $\gamma \simeq 2.7$ for the full day.

nodes, are weighted by the number of people moving in and out of the locations during a typical day. The clustering coefficient, measuring the local connectedness of the graph, scales as k^{-1} (k is the degree of the node) for sufficiently large k. This scaling is consistent with that obtained from models that postulate underlying hierarhical structures (few nodes get most of the action). The out-degree distribution in log-log scale is relatively constant for small k but exhibits power law decay afterwards ($P(k) \propto k^{-\gamma}$). The distribution of daily total out-traffic between nodes in log-log scale is flat for small k but exhibits power law decay afterwards. The distribution of the daily out-traffic of individuals between nodes scales as a power law for all k (degree).

The observed power law distribution in the out-traffic (edge weights) is therefore, supportive of the theoretical analysis of Yook *et al.* [33] who built weighted scale-free (WSF) dynamic networks and proved that the distribution of the total weight per node (total out-traffic in our network) is a power law where the weights are exponentially distributed.

There have been limited attempts to identify at least some characteristics of the joint distributions of network properties. The fact that daily out-degree and total out-traffic data are

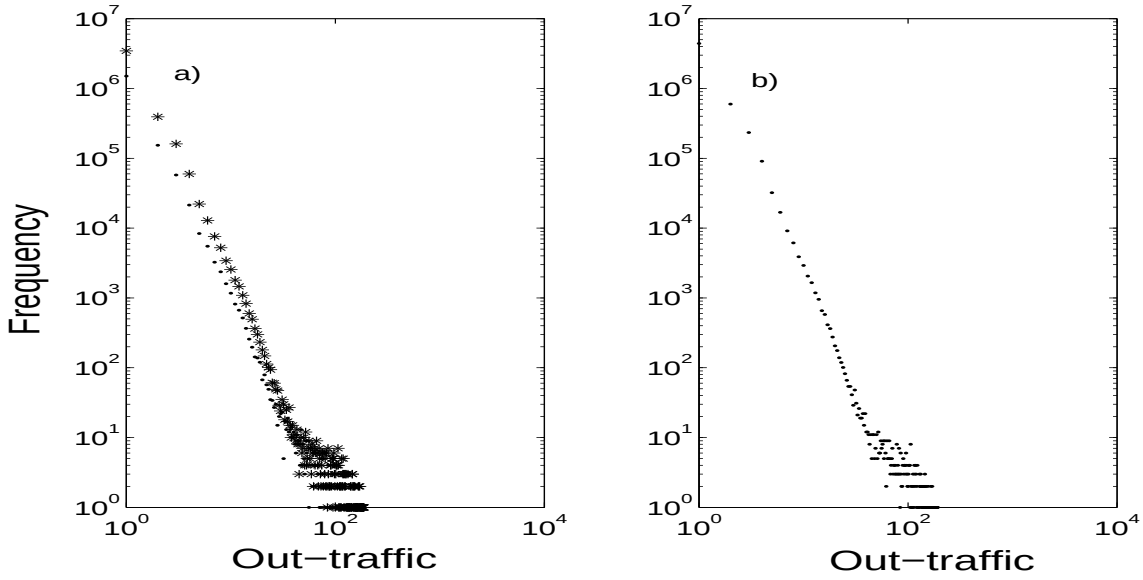

Figure 7: The out-traffic plots of the location-based network of the city of Portland has a power law distribution $\propto k^{-\gamma}$ with (a) $\gamma \approx 3.56$ (morning), $\gamma \approx 3.74$ (afternoon), and (b) $\gamma \approx 3.76$ (full day). Hence a few connections have high traffic but most connections have low traffic.

highly correlated is consistent again with the results obtained from models that assume an underlying hierarhical structure (few nodes have most of the connections and get most of the traffic (weight)). The Portland network exhibits a strong linear correlation between out-degree and total out-traffic on a log-log scale. We use this time series data to look at the network "dynamics" as the activity in the network increases, the size of the maximal connected component exhibits threshold behavior, that is, a "giant" connected component, suddenly emerges. The study of superimposed processes on networks such as those associated with the potential deliberate release of biological agents needs to take into account the fact that traffic is not constant. Planning, for example, for worst-case scenarios requires knowledge of edge-traffic, in order to characterize the temporal dynamics of the largest connected network components [34].

6 Acknowledgements

The authors thank Pieter Swart, Leon Arriola, and Albert László Barabási for interesting and helpful discussions. This research was supported by the Department of Energy under contracts W-7405-ENG-36 and the National Infrastructure Simulation and Analysis Center (NISAC).

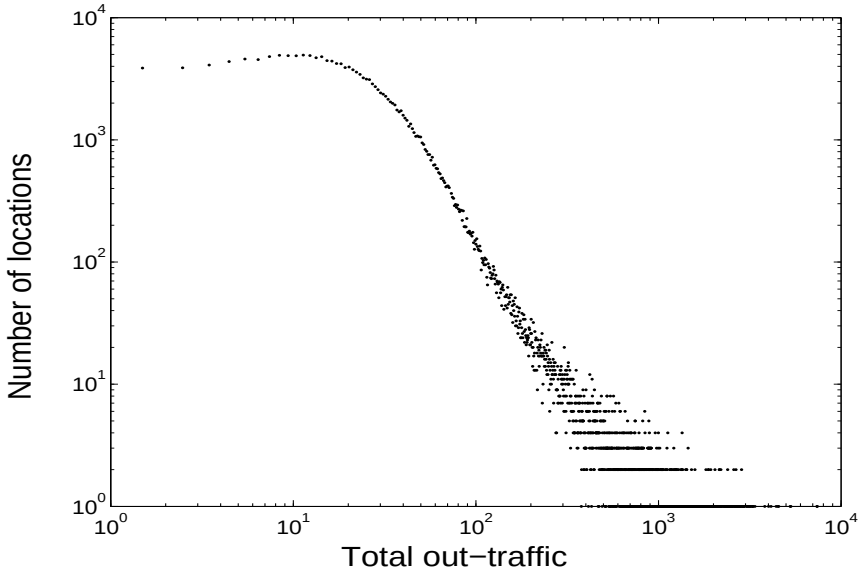

Figure 8: Distribution of the total traffic for the location-based network of the city of Portland. There are approximately the same number of locations (nodes) with small total out-traffic. The number of locations where more than 30 people (approximately) leave each day decays as a power law with $\gamma \simeq 2.74$.

References

[1] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. 98, pp. 404-409 (2001).

[2] M. E. J. Newman, *Who is the best connected scientist? A study of scientific coauthorship networks*, Physical Review E 64 (2001) 016131; Phys.Rev. E64 (2001) 016132.

[3] A.-L. Barabsi, H. Jeong, R. Ravasz, Z. Nda, T. Vicsek,and A. Schubert, *On the topology of the scientific collaboration networks*, Physica A 311, 590-614 (2002).

[4] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási, *The large-scale organization of metabolic networks*, Nature 407, 651-654 (2000).

[5] H. Jeong, S. Mason, A.-L. Barabási, and Z.-N. Oltvai, *Lethality and centrality in protein networks*, Nature 411, 41-42 (2001).

[6] M. Faloutsos, P. Faloutsos, C. Faloutsos, *On Power-Law Relationships of the Internet topology*, SGCOMM (1999).

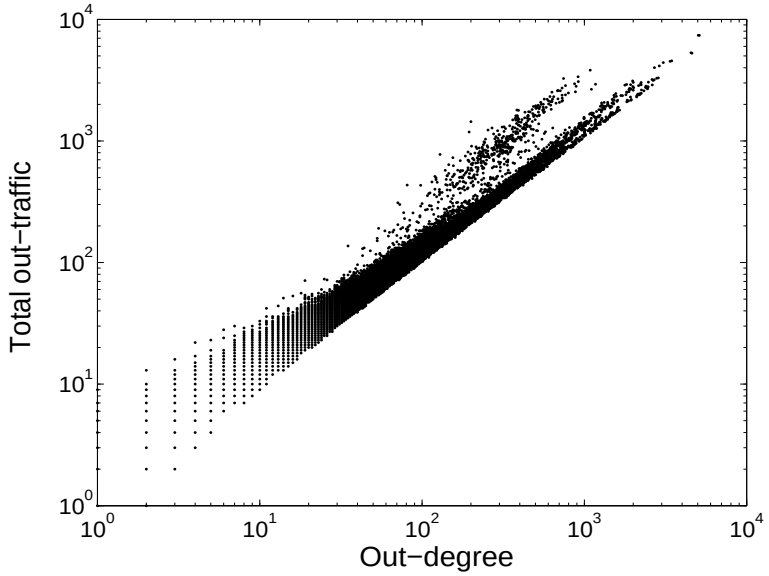

Figure 9: The out-degree and the total out-traffic are highly correlated with correlation coefficient $\rho = 0.94$ on a log-log scale. Most (95%) of the locations have fewer than 100 people leaving during the day.

[7] R. Albert, H. Jeong, and A.-L. Barabási, *Diameter of the World Wide Web*, Nature 401, 130-131 (1999).

[8] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew S. Tomkins, Eli UpfalProc (200). 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS.

[9] D. J. Watts and S. H. Strogatz, *Collective Dynamics of Small-World Networks*, Nature, 363:202-204 (1998).

[10] S.H. Strogatz, *Exploring Complex Networks*, Nature 410, 268-276 (2001).

[11] B. Bollobás, *Random Graphs*, Academic, London (1985).

[12] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai,and A.-L. Barabási, *Hierarchical Organization of Modularity in Metabolic Networks*. Science 297, 1551-1555 (2002).

[13] Erzsébet Ravasz and Albert-László Barabási, *Hierarchical organization in complex networks*, Physical Review E (in press).

[14] C. L. B. . et al. TRANSIMS: Transportation Analysis Simulation System. Technical Report LA-UR-00-1725, Los Alamos National Laboratory, Unclassified Report, 2001.

[15] C. Barrett, K. Bisset, R. Jacob, G. Konjevod, and M. Marathe. An Experimental Analysis of a Routing Algorithm for Realistic Transportation Networks. *to appear in European Symposium on Algorithms (ESA)*, Sept. 2002. Technical Report No. LA-UR-02-2427, Los Alamos National Laboratory.

[16] C. Barrett, R. Jacob, and M. Marathe. Formal Language Constrained Path Problems. *SIAM J. Computing*, 30(3):809–837, 2001.

[17] R. Jacob, M. Marathe, and K. Nagel. A Computational Study of Routing Algorithms for Realistic Transportation Networks. *ACM J. Experimental Algorithmics*, 4:Article 6, 1999. http://www.jea.acm.org/1999/JacobRouting/.

[18] Albert-László Barabási, Réka Albert, Hawoong Jeong, *Mean-field theory for scale-free random networks*, PHYSICA A 272 (1999) 173-87.

[19] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, *Classes of small-world networks*, Proc. Natl. Acad. Sci. (2000).

[20] P. Erdos and A. Renyi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[21] D. Peterson, L. Gatewood, Z. Zhuo, J. J. Yang, S. Seaholm, and E. Ackerman. Simulation of stochastic micropopulation models. *Computers in Biology and Medicine*, 23(3):199–210, 1993.

[22] D. J. Watts, *Small Worlds: The dynamics of networks between order and randomness*, Princeton University Press (1999).

[23] R. Sedgewick, *Algorithms*, Addison-Wesley (1988).

[24] P. Grassberger, *On the critical behavior of the general epidemic process and dynamical percolation*, Math. Biosc. 63, 157-172 (1983).

[25] M. E. J. Newman, *The spread of epidemic disease on networks*, Phys. Rev. E 66, 016128 (2002).

[26] M. E. J. Newman, *Percolation and epidemics in a two-dimensional small world*, Phys. Rev. Lett. 66, 021904 (2002).

[27] C. Moore and M. E. J. Newman,*Epidemics and percolation in small-world networks*, Phys. Rev. Lett. 61, 5678-5682 (2000).

[28] R. Pastor-Satorras and A. Vespignani, *Epidemic dynamics and endemic states in complex networks*, Phys. Rev. Lett. 63, 066117 (2001).

[29] R. Pastor-Satorras and A. Vespignani, *Immunization of complex networks*, Phys. Rev. Lett. 65, 036104 (2002).

[30] R. M. May and A. L. Lloyd, *Infection dynamics on scale-free networks*, Phys. Rev. Lett. 64, 066112 (2001).

[31] Z. Dezso and A.-L. Barabási, *Halting viruses in scale-free networks*, Rev. Lett. 65, 055103 (2002).

[32] V. M. Eguíluz and K. Klemrn, *Epidemic threshold in structured scale-free networks*, Phys. Rev. Lett. 89, 108701 (2002).

[33] S.H. Yook, H. Jeong, A.-L. Barabási and Y. Tu, *Weighted Evolving Networks*, Physical Review 86, 5835-5838 (2001).

[34] G. Chowell and C. Castillo-Chavez, *Worst-Case Scenarios and Epidemics*, Los Alamos Unclassified Report LA-UR-03-1409 (2003).

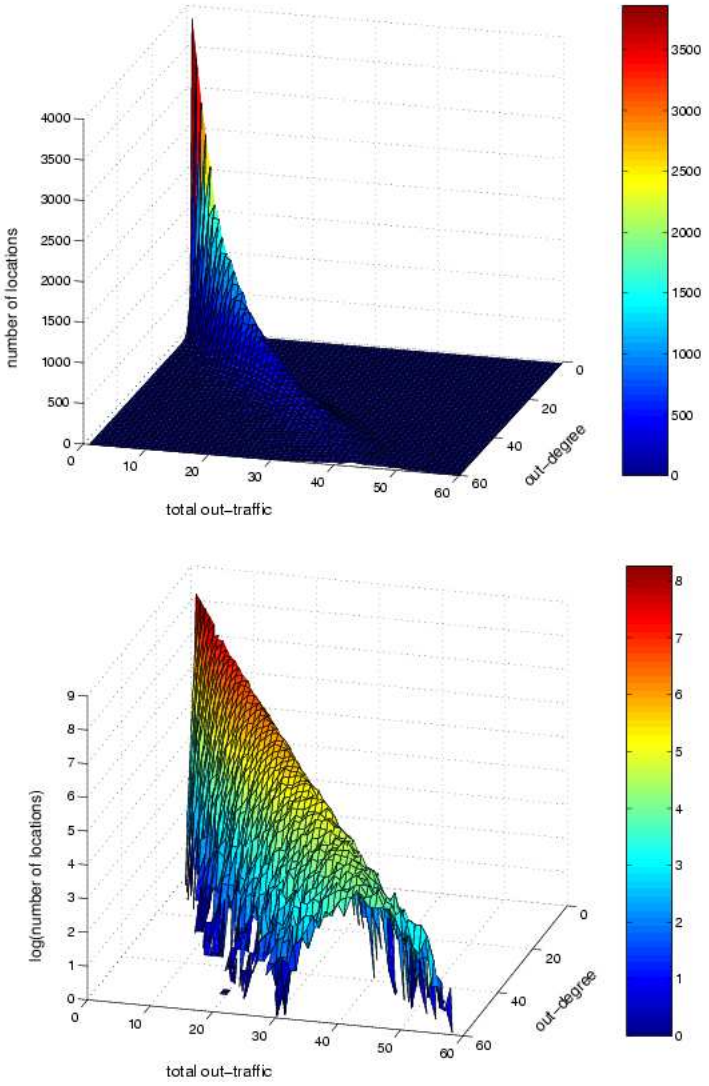

Figure 10: (a)Joint distribution $F(k, v)$ plot (b)log_e density of $F(k, v)$ plot between the out-degree k and the total out-traffic v in the location-based network of the city of Portland.